

September 9, 2003

MEMORANDUM

To: NIH Investigators

From: Gunta J. Lidars
Director, Office of Research and Project Administration

Marjorie Hunter
Director, Medical Center Office of Technology Transfer

RE: NIH Application Requirement for Data Sharing Plans

As previously disseminated in March, the NIH will be requiring a data sharing plan with all applications exceeding \$500,000 in direct costs in any single year (or if it is a special requirement in the program announcement) beginning with the October 1, 2003 deadline. This notice serves as a reminder of this requirement for the **timely release and sharing of final research data from NIH-supported studies** and also provides some suggestions for drafting such plans for NIH applications.

NIH does provide investigators with examples of acceptable methods of data sharing on its web site at http://grants1.nih.gov/grants/policy/data_sharing/index.htm, specifically in its "Data Sharing Workbook". Of specific importance to the investigator in formulating a data sharing plan:

- Final research data for the purposes of developing the plan are defined as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings. Final research data do not include laboratory notebooks, partial data sets, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as gels or laboratory specimens."
- Data sharing for a small laboratory-based project may be best accomplished by making raw data available in publications (e.g., via appendices) or under the auspices of the investigator. For larger studies as perhaps evidenced by applications over the \$500,000 threshold, providing data in publications is probably not the best vehicle for sharing. Therefore, other means for sharing must be made available such as through data archives, web sites or enclaves.
- The type of plan that an investigator will develop will ultimately be discipline and project specific, and should be tailored to each individual situation.
- Applicants may request funds for data sharing and archiving in grant applications.

- Data sharing plans should be provided in the form of a brief paragraph (longer if necessary) immediately following the Research Plan section of the NIH 398 Application.

Internet based data sharing is particularly amenable to large multi-site or multi-investigator projects. The University of Rochester now has one additional avenue available for data set sharing that may be available to NIH investigators affected by this requirement. DSpace (<http://www.dspace.org/>) is a repository system that enables institutions to capture, preserve and distribute investigator controlled and selected research data. DSpace currently is available for NIH research projects that require less than 100MB of storage capacity. For additional information about DSpace contact Julia Sollenberger, Director, Health Science Libraries and Technologies, x5-5194 or Julia_Sollenberger@urmc.rochester.edu.

Attached are several examples of data sharing plans that may be appropriate for certain research projects. Note that these have been provided as examples only, and the Office(s) of Technology Transfer are available to assist with specific cases. (Medical Center Office of Technology Transfer (x3-3743) or the University Office of Technology Transfer (x5-3998))

In addition, data sharing plans should take into account proprietary information of the University and third parties, protection of human subjects information and Protected Health Information and other situations where data sharing may not be appropriate or allowed.

Attachment (Examples of Data Sharing Plans)

[h:/letters/gjl/nih data sharing plans dspace](h:/letters/gjl/nih_data_sharing_plans_dspace)

pc: M. Coburn
C. Phelps
H. Federoff
T. Pearson
P. Slattery
D. Krusch

Examples of Data Sharing Plans for NIH Applications

As noted in the NIH Policy, “the precise content of the data-sharing plan will vary, depending on the data being collected and how the investigator is planning to share the data. Applicants who are planning to share data may wish to describe briefly the expected schedule for data sharing, the format of the final dataset, the documentation to be provided, whether or not any analytic tools also will be provided, whether or not a data-sharing agreement will be required and, if so, a brief description of such an agreement (including the criteria for deciding who can receive the data and whether or not any conditions will be placed on their use), and the mode of data sharing (e.g., under their own auspices by mailing a disk or posting data on their institutional or personal website, through a data archive or enclave). Investigators choosing to share under their own auspices may wish to enter into a data-sharing agreement.”

Please note the following two examples are illustrative, and may not be appropriate or suitable for data sharing depending on project and/or discipline.

1) Example from the NIH Policy:

The proposed research will include data from approximately 500 subjects being screened for three bacterial sexually transmitted diseases (STDs) at an inner city STD clinic. The final dataset will include self-reported demographic and behavioral data from interviews with the subjects and laboratory data from urine specimens provided. Because the STDs being studied are reportable diseases, we will be collecting identifying information. Even though the final dataset will be stripped of identifiers prior to release for sharing, we believe that there remains the possibility of deductive disclosure of subjects with unusual characteristics. Thus, we will make the data and associated documentation available to users only under a data-sharing agreement that provides for: (1) a commitment to using the data only for research purposes and not to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.

2) Example for array data sharing proposal:

The generation and analysis of microarray data is a time consuming and costly proposition. Tremendous resources in the form of reagents and personnel are needed to ensure that the data acquired using microarray-based technologies is robust and reproducible. In addition, the ability to share both raw and processed data is critical for success at the level of the laboratory, institution/region and nationally sponsored programs. To this end, we have established a comprehensive program that will allow us to effectively share our data with the scientific community at many levels while efficiently incorporating data generated by others into our analytical scheme. We will do this in three ways:

(A) Laboratory/Institutional: We have deployed a piece of technology that will allow us to effectively organize, archive and analyze all of our microarray data at a single source. We are using Iobion's GeneTraffic MULTI package to accomplish this task. The laboratory information management (LIMS) piece of this package will allow us to coordinate all microarray data being generated in our program as well as other investigators at University of Rochester. Access to raw and processed data is performed through our local area network (LAN) without sacrificing data security. Accounts will be generated for all personnel in the laboratory in order to facilitate gene expression analysis using approaches described in our application.

(B) Regional: A consortium has been launched whose mission is to coordinate all microarray-based activities at participating medical schools and hospitals. We are proud to be a part of this consortium effort and have already benefited from having access to regional and national expertise in the microarray field. Being involved in this consortium effort lets us leverage our biological expertise against a wealth of information in the genomics field without duplicating technology infrastructure while giving us access to cutting edge advancements in protocol development and implementation. To this end, the consortium created and implemented the distributed data sharing infrastructure throughout the region by helping participating institutions (including University of Rochester) acquire and implement the Iobion GeneTraffic Servers and software. We have access to all of the servers in the network (through the internet) which has already helped facilitate data sharing and management across the region. In addition, we are using similar analysis and visualization tools that are helping standardize the biological assessment of our array data. As this structure continues to develop the transport of raw and analyzed data will be seamless across participating institutions. Lastly, a centralized data repository is being constructed allowing participating institutions the ability to communicate with national data repositories in an efficient manner helping expand the "power" of analysis at individual institutions.

(C) Federal: It is equally important that we contribute to NIH driven data repositories such as GEO. Due the fact that the GeneTraffic data scheme is MIAME compliant we will be prepared to upload any or all of our data to GEO or other national repositories as a part of our discovery process. By incorporating the Minimal Information About Microarray Experiment (MIAME) into our data structure we will be able to communicate with investigators outside of the consortium using these set of standards to better compare and contrast results obtained through our microarray data analysis.

In short, our implementation of specific standard operating procedures and involvement with the consortium makes data generated from this proposal available on a global level and will help strengthen our own discovery process through efficient data sharing mechanisms.

3) *Data sharing using University of Rochester's DSpace institutional repository system*

We will be sharing our research data via a digital data archive. DSpace is a repository system, administered by the University of Rochester Libraries, that enables the capture, preservation and distribution of the intellectual output of the faculty, researchers, and staff of the University of Rochester. The works deposited in the repository are scholarly, educational or research oriented and must be in digital format. Examples of items that DSpace can accommodate are: articles, preprints, working papers, technical reports, conference papers, books, theses, **data sets**, computer programs, and visual simulations and models.

DSpace is organized into “communities” and “collections,” each of which retains its identity within the repository. Customization for communities and collections allows for flexibility in determining policies and workflow. Authors depositing content must be willing and able to grant the University of Rochester the right to preserve and distribute the work via DSpace, although the author retains copyright for all works submitted. Authors do have the right to specify the level of access to their own content (at the “collection” level and the individual item level) to world wide, UR-only, or to specific individuals or groups.

DSpace provides long-term physical storage and management of digital items in a secure, professionally managed repository including standard operating procedures such as backup, refreshing media, and disaster recovery. The system assigns a persistent identifier to each contributed item to ensure its retrievability. DSpace also provides a mechanism for advising content contributors of the preservation support levels they can expect for the files they submit.

The DSpace submission process allows for the description of each item using a specified metadata schema. These descriptions are entered into a relational database, which is then used by the DSpace search engine to retrieve items. The system also opens up the metadata to broader search engines that search multiple repositories and “harvest” metadata from many sources in a distributed and decentralized electronic information environment. Authors themselves decide whether or not they want their content to be searchable and retrievable outside of the institution. Access can be limited by individual or group.

DSpace provides a convenient, stable, reliable, and cost-effective means for storing, preserving, and sharing digital content. Data sets deposited in DSpace are both protected and available.